DOCUMENT RESUME

ED 208 038 TM 810 735

AUTHOR Cook, Linda L.; And Others
TITLE IRT Equating: A Flexible Alternative to Conventional Methods for Solving Practical Testing Problems.
PUB DATE Apr 81
NOTE 56p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS College Entrance Examinations; Comparative Analysis; *Equated Scores; *Feasibility Studies; *Latent Trait Theory; *Mathematical Models; Methods
IDENTIFIERS Equipercentile Equating; Frequency Estimation Equipercentile Equating; Linear Equating Method; *National Merit Scholarship Qualifying Test; *Preliminary Scholastic Aptitude Test

ABSTRACT

The purposes of this study are: (1) to compare the results of linear, equipercentile, frequency estimation equipercentile and item response theory (IRT) true formula score equating; and (2) to investigate the feasibility of using IRT methods to equate new forms of the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) to each other directly. Equating samples for all methods, except the frequency estimation approach, contained approximately 2,000 randomly-selected cases from data obtained at the regular administrations of each of the old and new forms. Larger samples were used for the frequency estimation approach. The most notable aspect of the results obtained from the comparison of the four methods was the marked agreement found among them. The results also indicated that it is feasible to use IRT methods to equate the two forms of the PSAT/NMSQT directly. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208038

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

IRT Equating: A Flexible Alternative to Conventional Methods for Solving Practical Testing Problems[1]

Linda L. Cook[2]
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. L. Cook

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Stephen B. Dunbar
University of Illinois

Daniel R. Eignor
Educational Testing Service

TM 810 735

[1] A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981.

[2] The authors names appear in alphabetical order. The authors fully acknowledge the advice provided by Gary L. Marco and Martha L. Stocking and the technical assistance provided by Karen Zeis.

IRT Equating: A Flexible Alternative to Conventional Methods for Solving Practical Testing Problems

Linda L. Cook
Educational Testing Service

Stephen B. Dunbar
University of Illinois

Daniel R. Eignor
Educational Testing Service

INTRODUCTION

In spite of extensive efforts on the part of test development personnel to ensure that multiple forms of a test are similar in content and difficulty, form to form differences tend to occur with regular frequency. This situation requires that some adjustment be made to the scores on different forms of the test before test results can be interpreted in a meaningful way. The extent to which such adjustments are free from statistical bias clearly affects the extent to which later substantive interpretations of test scores are bias free. Thus, to ensure fairness to examinees taking different forms of a test and competing for the same positions, accuracy in this process, hereafter referred to as equating, is essential.

The current thrust of research devoted to the practical applications of item response theory (IRT) has generated an active interest in score equating. While this interest is anything but new, it is one which calls attention to the underlying assumptions of the equating methods used by many large scale testing programs. In an effort to understand more about the effects of equating on the integrity of score scales, this study assesses the relative agreement of four methods: (1) linear; (2) equipercentile; (3) frequency estimation equipercentile; and, (4) IRT estimated true formula score equating.

In addition, a unique application of IRT methods is presented which demonstrates their flexibility in solving equating problems not amenable to traditional methods. The data used for the study came from two recent administrations of the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT), a test which is developed and administered by the College Board Admissions Testing Program.

BACKGROUND AND PURPOSE OF THE STUDY

A variety of definitions of equated scores have appeared in the literature, the most general and perhaps restrictive being that of Lord (1977), in which he argues that "...transformed scores, y*, and raw scores, x, can be called equated if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y." In principle, Lord's definition subsumes equating of both non-parallel and parallel forms; but, as he explains, one would not expect these requirements to be met unless strictly parallel forms were being used. This is because tests (forms) that are not strictly parallel will differ in level of difficulty. Forms that differ in difficulty cannot, because of their true score relationship, be equally reliable. It is certainly not a matter of indifference to an examinee, particularly a high ability examinee, whether he/she takes one form of a test that is less reliable than a second form. A somewhat relaxed way of characterizing the notion of equivalent scores (Angoff, 1971) is to say that scores on two test forms may be considered equivalent if they have identical frequency distributions for some population of examinees.

Whatever definition of equivalent scores is adopted, two considerations are relevant to obtaining them, a design for data collection and a statistical

model for transforming raw scores to a common scale. Angoff (1971) provides a comprehensive review of equating designs and their concomitant assumptions and transformation procedures. Designs range from the simple single group (one group, two test forms), to the random groups (two randomly equivalent groups, two test forms), to the more complicated anchor test (two not necessarily equivalent groups, two test forms and one "anchor" test of common items taken by both groups). The design used to equate the PSAT/NMSQT is a complex version of the basic anchor test design.

Standard practice in equating new forms of the PSAT/NMSQT is to equate each new form of the test to two old forms of the Scholastic Aptitude Test (SAT) through separate sets of common items. One can imagine <u>each</u> of the two new PSAT/NMSQT forms produced annually as being composed of three sets of items: (1) items unique to that form; (2) items in common with one old SAT form; and (3) items in common with a second old SAT form. It is important to note that both new forms (Form 1 and Form 2) of the PSAT/NMSQT share items in common with the same two old SAT forms. However, there exists no item overlap between the two new forms, i.e., each new form is equated back to the same two old SAT forms but through <u>different</u> sets of common items.

Final scaled scores are determined separately for each new form as follows: (1) the results of the PSAT/NMSQT Form 1 linear equating to the <u>first</u> SAT old form and the results obtained from the PSAT/NMSQT Form 1 linear equating to the <u>second</u> SAT old form are bisected, if the new to old forms relationships are judged to be linear; (2) the results of the PSAT/NMSQT Form 1 equipercentile or frequency estimation equipercentile equating to the <u>first</u> old SAT form and the results of the PSAT/NMSQT Form 1 equipercentile or frequency estimation equipercentile equating to the <u>second</u> old SAT form are

averaged, if the new to old forms relationships are judged to be curvilinear. This process is repeated for the PSAT/NMSQT Form 2 equating. It should be noted that the two PSAT/NMSQT new forms are related (equated to each other) only through their relationship to the same two old SAT forms. It is not possible to equate the new forms directly by traditional methods because they contain no common items and are given to non-randomly equivalent groups.

Constraints imposed by the PSAT/NMSQT data collection design present several potential problems for the equating process. First, several not necessarily equivalent groups are represented in the design. The two PSAT/NMSQT equating samples (selected from the Form 1 and Form 2 populations) are potentially non-randomly equivalent because of self selection of testing date. Moreover, the two SAT equating samples (selected from the first and second old form populations) are non-randomly equivalent with respect to the PSAT/NMSQT groups to the extent that they differ in level of ability. A second potential problem stems from differences between the PSAT/NMSQT and SAT in length, reliability, and level of difficulty.

One might reasonably expect IRT methods to offer several advantages over traditional methods, at least as far as the PSAT/NMSQT design is concerned. First of all, according to Lord (1975), "In theory ICC (IRT) methods are capable of estimating the equipercentile line of relation between raw scores when two tests to be equated are not parallel, are given to non-equivalent groups, and everyone takes an anchor test. Strictly speaking, no other method known to the writer can accomplish this." Second, as explained in detail at a later point in this paper, it is possible to employ IRT methods to equate new forms of the PSAT/NMSQT directly to one another even through they contain no common items and are given to non-randomly equivalent groups.

The purpose of this study, therefore, is twofold: (1) to compare the results of linear, equipercentile, frequency estimation equipercentile and IRT true formula score equating under the constraints of the PSAT/NMSQT design; and (2) to investigate the feasibility of using IRT methods to equate new forms of the PSAT/NMSQT to each other directly. Results from the first part of the study will provide some indication of the relative agreement of the four methods, whereas those of the second will illustrate the flexibility of IRT approaches in solving a heretofore intractable testing problem.

RELATED RESEARCH

A number of researchers have recently investigated the relative performance of score equating procedures applied to different equating designs in horizontal and vertical equating situations. While it is fair to say that, on a very general level, a certain degree of consensus exists as to which procedures yield the most accurate results, the differences between the findings of these studies, particularly those related to the stability of results, is a cause for concern. Slinde and Linn (1977, 1978, 1979) investigated in an indirect fashion the problem of vertical equating of two forms designed for populations at different levels of ability. Their results suggested that linear, equipercentile and IRT equating employing the one-parameter logistic model may have limitations for the process of vertical equating. This was especially true when the differences between test difficulty and between ability levels of equating samples were most pronounced. Their studies imply that an IRT approach based on the more complex three-parameter logistic model might provide more useful results for vertical equating situations.

Marco, Petersen, and Stewart (1979) presented perhaps the most compre-

hensive empirical study of equating techniques yet to appear. For designs similar to the PSAT/NMSQT design, they found problems with traditional methods similar to those found in the Slinde and Linn studies. In particular, when tests differing in difficulty were given to non-randomly equivalent groups and equated using an anchor test design, traditional procedures appeared to break down. In spite of the presence of possible criterion bias confounding some of their results, the authors suggested that the three-parameter logistic model would yield the most acceptable results under unusual or extreme design constraints. However, Marco et al found, as did Slinde and Linn, that the degree of dissimilarity between groups and test forms were both relevant. When these factors were moderate, traditional methods, both linear and equipercentile, yielded adequate equatings.

A comparison of the stability of results obtained from traditional and IRT procedures was made by Kolen (1981), who used a cross-validation group to establish a criterion for the evaluation of seven IRT methods and two traditional methods (linear and equipercentile). Kolen had some difficulty evaluating the results obtained from application of the three-parameter logistic model to equate new Level I tests (vocabulary and quantitative thinking tests administered to 9th and 10th graders) and new Level II tests (tests of the same skills administered to 11th and 12th graders) to old tests of vocabulary and quantitative thinking that consisted of one level, administered to grades 9-12. He found that "Although the three-parameter estimated observed score method tended to produce the most stable cross-validation results at Level I of the tests, the results were of only moderate accuracy at Level II. The three-parameter estimated true score equivalents method tended to produce the most stable cross-validation results at Level II but results of moderate

stability at Level I." Kolen noted that these results may be partially explained by difficulty in estimating "c" parameters and by problems related to the transformation of estimated true scores below the chance score level; he concluded that there may be some doubt regarding the applicability of the three-parameter model in all equating contexts. A somewhat surprising degree of stability was found to hold for conventional equipercentile procedures, sufficient to lead Kolen to conclude that they may be the most viable in practice for designs which involve equating tests differing only moderately in difficulty.

More recently, Petersen, Cook, and Stocking (1981) investigated the relative accuracy of conventional versus IRT equating methods using scale drift as the criterion for evaluating these methods. The authors used a chain of six verbal and six mathematical forms of the SAT, each form linked to the preceeding and following form by an anchor test. The design allowed them to evaluate the affect of equating a test to itself via five intervening forms. The equating methods investigated were four traditional methods (three linear and one equipercentile) and a three-parameter logistic IRT method that employed three procedures for linking parameter estimates so that estimated true formula scores could be equated. The three linking procedures were: (1) concurrent calibration (the method used in this study); (2) fixed b's; and (3) equated b's. Results of the study indicate that all three IRT linking methods were superior to the traditional methods for the verbal portion of the test. For the mathematical portion, the IRT fixed b's and equated b's methods did not produce results as satisfactory as those obtained from two of the three linear methods. However, the IRT concurrent method produced results comparable to these two methods. Equipercentile equating yielded fairly unsatisfactory

results for both the verbal and mathematical sections. It should be noted that the study involved tests similar in level of difficulty which were given to groups of examinees that did not differ greatly in their level of ability; a situation in which one would expect traditional linear methods to work well.

If anything, an in-depth look at previous research comparing various equating procedures leaves the practitioner a little bit bothered. On the one hand, IRT approaches, especially those using the three-parameter logistic model, appear to provide the most accurate results and hence seem appropriate from an empirical perspective as well as a theoretical one. On the other hand, there is some question regarding their stability, although the comparatively small amount of scale drift associated with the IRT concurrent calibration design found by Petersen et al (1981) is evidence in support of their application to parallel forms of aptitude tests administered to groups that are similar in ability. In addition, it is important to note that the studies reviewed indicate that at present the effects of differential reliability and difficulty of test forms and the effect of the non-randomness of examinee samples do not appear to be completely understood.

DATA COLLECTION AND METHOD

The data for this study came from two recent administrations of the PSAT/NMSQT, a test which is developed and administered by the College Board Admissions Testing Program. Also used were data from two forms of the SAT, developed and administered by the same organization. Both the PSAT/NMSQT and SAT are multiple choice tests. The tests differ in length and difficulty, the PSAT/NMSQT being composed of 65 verbal and 50 mathematical items and the SAT

of 85 verbal and 60 mathematical items. The PSAT/NMSQT consists of two 50-minute sections. The verbal section contains only 5-choice items; the mathematical section contains a mixture of 4- and 5-choice items. Raw scores obtained on the PSAT/NMSQT are most typically transformed to scaled scores on the College Board 200 to 800 scale via the linear equating method described on page 3. For score reporting purposes, the final digit of the score is dropped and scores are reported on a 20 to 80 scale. PSAT/NMSQT raw scores are actually formula scores generated from number right scores using a correction for guessing formula. Raw scores are computed by the formula $R - kW$, where R is the number of correct responses, W is the number of incorrect responses and $k = 1/n-1$, n being the number of choices per item. Both the verbal and mathematical sections of the test were used for this study.

The SAT consists of six 30-minute sections: two verbal sections, two mathematical sections, one Test of Standard Written English (TSWE) and one experimental section containing an equating test or pretest. The two verbal sections, one mathematical section and the TSWE contain 5-choice items; the other mathematical section contains a mixture of 4- and 5- choice items. Raw scores on the SAT are also typically transformed to scaled scores on the College Board 200 to 800 scale by linear equating methods. This scale is retained for score reporting. SAT raw scores are formula scores incorporating the correction for guessing procedure previously described. Only the two verbal and two mathematical sections of the test were used for the study.

Figure 1 illustrates the equating design employed for the first part of the study, which involves assessing the relative agreement of IRT and traditional methods. PSAT/NMSQT Form 1 and Form 2 are alternate forms of the PSAT/NMSQT, each containing a subset of items in common with each of the SAT old forms

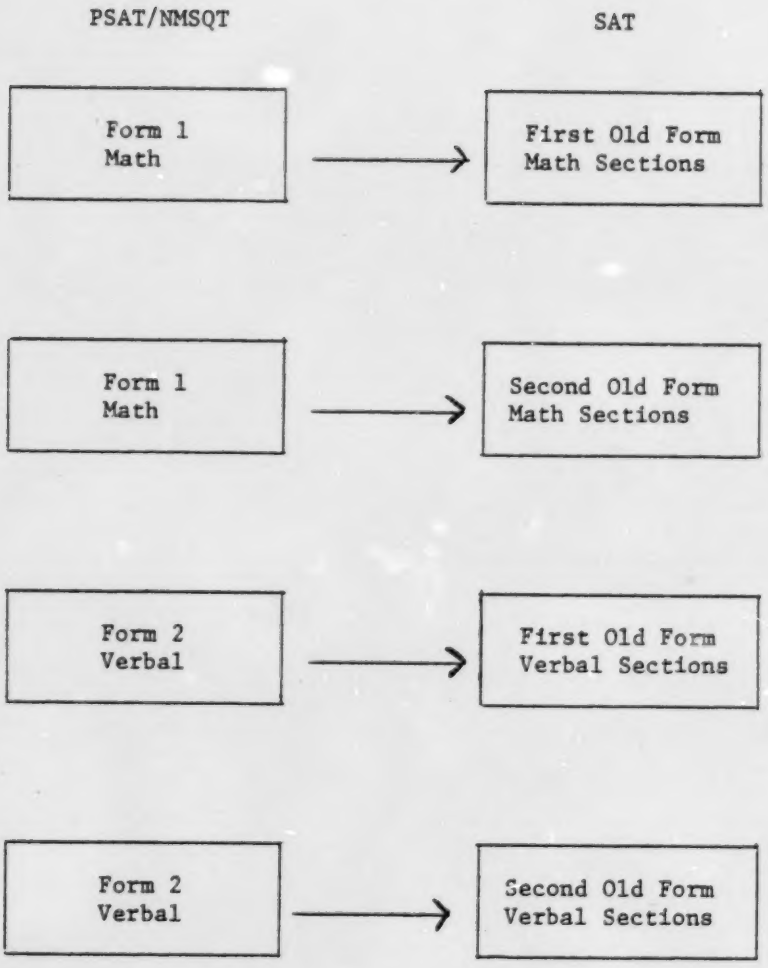


Figure 1: Schematic Diagram of Design Used in Study for Equating PSAT/NMSQT Form 1 Math and Form 2 Verbal to SAT First and Second Old Forms.

(hereafter designated SAT First Old Form and SAT Second Old Form.) Only the mathematical section of PSAT/NMSQT Form 1 and the verbal section of PSAT/NMSQT Form 2 were examined for agreement across methods.

Equating samples for all methods, except the frequency estimation approach, contained approximately 2,000 randomly-selected cases from data obtained at the regular administrations of each of the old and new forms shown in Figure 1. A total of eight random samples, two each for PSAT/NMSQT Form 1 and PSAT/NMSQT Form 2, SAT First Old Form and SAT Second Old Form, were used in the study. Since sample sizes of 2,000 are not likely to yield stable estimates for the frequency estimation procedure, separate, larger samples (approximately 9,000 cases for each PSAT/NMSQT sample and 5,000 cases for each SAT sample) were used for this approach. Tests for differences between the frequency estimation equating samples and those drawn for the other methods indicated that no significant differences existed at the .05 level.

As mentioned previously, four separate PSAT/NMSQT to SAT anchor test equatings (two verbal and two mathematical) were repeated for each of the four methods of interest; linear, equipercentile, frequency estimation equipercentile, and IRT true formula score. Each of these methods is described in greater detail below. Appendix A provides additional information regarding conversion procedures.

The basis for the linear conversions under consideration is that scores on two test forms are equivalent if they correspond to the same number of standard deviations from the mean in some group of examinees. The linear methods used were either the Tucker or Levine models (cf. Angoff, 1971). Both of these models assume that scores on the relevant selection attribute (the attribute on which the equating samples vary) are collinear with the

scores on the anchor test.

Each of the equipercentile models maintains that scores on two test forms are equivalent if they correspond to the same percentile rank in some group of examinees. The ordinary equipercentile procedure involves equating scores on each test form to the anchor test separately within each group. Scores on the two forms to be equated are then said to be equivalent if they correspond to the same score on the anchor test. In contrast, the frequency estimation equipercentile method estimates the frequency distributions of scores on the two forms of interest for a hypothetical combined group of examinees (students who took the new form and students who took the old form). Again, scores on the two forms are said to be equivalent if their corresponding percentile ranks are the same.

Finally, IRT equating models characterize equivalent scores on two test forms as those scores which correspond to the same estimated level of the latent trait, ability, or skill underlying both tests. Item response theory assumes that a mathematical function relates the probability of a correct response on an item to an examinee's ability (Lord, 1980). As previously mentioned, the mathematical function (IRT model) employed in this study was the three-parameter logistic model. The model states that the probability of a correct response to item i ($P_i(\theta)$), is given by:

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{1.7a_i(\theta-b_i)}}, \quad (i=1, 2, \ldots, n), \qquad (1)$$

where, a_i, b_i, and c_i are three parameters describing the item and θ represents the ability level of an examinee.

The item parameters and examinee abilities for the study were estimated

using the program LOGIST (Wood and Lord, 1976; Wood et al, 1976). The estimates are obtained by a (modified) maximum likelihood procedure which has been adapted to accommodate omitted items (Lord, 1974).

Although a variety of equating techniques exist once an IRT model has been chosen, only estimated true formula score equating (Lord, 1980, Chapter 13) was used for this study. Estimated true formula scores $\hat{\xi}$ and $\hat{\eta}$ on two tests measuring the same ability, θ, are related by the equations,

$$\hat{\xi} = \sum_{i=1}^{n} \hat{P}_i(\theta) - \left[\sum_{i=1}^{n} \hat{Q}_i(\theta)\right]/A-1 \qquad (2)$$

$$\hat{\eta} = \sum_{j=1}^{m} \hat{P}_j(\theta) - \left[\sum_{j=1}^{m} \hat{Q}_j(\theta)\right]/A-1 \qquad (3)$$

where, A is the number of choices per item, $\hat{P}_i(\theta)$, and $\hat{P}_j(\theta)$, represent the probability of a correct response for items i and j as they appear in the two forms to be equated and $\hat{Q}_i(\theta)$, $\hat{Q}_j(\theta)$ equal $1 - \hat{P}_i(\theta)$ and $1 - \hat{P}_j(\theta)$, respectively. Using expressions 2 and 3, it is possible to find an estimated true formula score $\hat{\xi}$ corresponding to an estimated true formula score $\hat{\eta}$ for any given θ.

Expressions 2 and 3 will not provide equated estimated true formula scores for scores on the two test forms of interest that fall below the chance score level. Several ways exist for determining the relationship in this region. Kolen (1981) used linear interpolation. The method that was used for this study involved estimating the mean and standard deviation of scores below the chance score level for the two forms of interest and using the estimated values to establish a linear relationship.

The means and standard deviations of below chance score level scores were

estimated using the following expressions:

$$M_x = \frac{A}{A-1} \sum_{i=1}^{n_x} c_i - \frac{n_x}{A-1} \tag{4}$$

$$S_x^2 = \left(\frac{A}{A-1}\right)^2 \left[\sum_{i=1}^{n_x} c_i - \sum_{i=1}^{n_x} c_i^2 \right] \tag{5}$$

where,

M_x = the mean of PSAT/NMSQT scores below chance level,

S_x^2 = the variance of PSAT/NMSQT scores below chance level,

A = the number of choices per item, and

c_i = the psuedo guessing parameter for item i .

Equations 4 and 5 were repeated to obtain M_y and S_y, the estimated mean and variance of below chance level scores for the SAT old form of interest. Linear parameters for equating PSAT/NMSQT scores below chance level to SAT scores below chance level were determined as follows:

$$A = \frac{S_y}{S_x} \tag{6}$$

$$B = M_y - AM_x \tag{7}$$

The linear parameters (A and B) are used to form the following expression:

$$\text{score (SAT)} = A\,[\text{score (PSAT/NMSQT)}] + B \tag{7}$$

The first part of the study involved the comparison of conventional linear and curvilinear methods with the IRT method. The item calibration plan for this part of the study is illustrated in Figure 2. Each of the four

First Equating

Group	PSAT/NMSQT Form 1 Math Items n=31	Common Items n=19	SAT First Old Form Math Items n=41
PSAT/NMSQT	X	X	Not Reached
SAT	Not Reached	X	X

Second Equating

Group	PSAT/NMSQT Form 1 Math Items n=30	Common Items n=20	SAT Second Old Form Math Items n=40
PSAT/NMSQT	X	X	Not Reached
SAT	Not Reached	X	X

Third Equating

Group	PSAT/NMSQT Form 2 Verbal Items n=42	Common Items n=23	SAT First Old Form Verbal Items n=62
PSAT/NMSQT	X	X	Not Reached
SAT	Not Reached	X	X

Fourth Equating

Group	PSAT/NMSQT Form 2 Verbal Items n=42	Common Items n=23	SAT Second Old Form Verbal Items n=62
PSAT/NMSQT	X	X	Not Reached
SAT	Not Reached	X	X

Figure 2: Calibration Plan for IRT Equatings Used for Comparison with Conventional Equatings

Each of the four boxes indicates a separate calibration run. Both new and old form samples contained 2000 cases. Crosses indicate items that examinee groups actually were exposed to.

separate boxes represents a single LOGIST run, yielding item and ability parameters on a common scale. Data for the separate runs are arranged such that each PSAT/NMSQT and SAT group is considered to have taken exactly the same test. For example, considering the first box in Figure 2, both groups are conceptualized as having taken a test composed of PSAT/NMSQT Form 1 Math items, common items, and SAT First Old Form Math items. Examinees are considered to simply not have reached those items to which they were not exposed. Ability estimates are thus based on a subset of "total" test items actually answered. The design permits true formula score equating of each PSAT/NMSQT - SAT pairing illustrated in Figure 2.

The calibration plan for the second part of the study, the direct equating of PSAT/NMSQT Form 1 Verbal scores to the PSAT/NMSQT Form 2 Verbal scores is shown in Figure 3. The entire matrix illustrated in Figure 3 represents a single LOGIST run. As before, each of the four groups is considered to have taken exactly the same test. This test is conceptualized as containing the eight components designated by the column headings in Figure 3. This plan permits direct equating of the PSAT/NMSQT Form 1 Verbal scores to the PSAT/NMSQT Form 2 Verbal scores even though the two sections contain no overlapping items. It also permits equating each of the PSAT/NMSQT Verbal scores separately to each of the SAT Verbal scores, thus allowing replication of the equatings carried out for these scores in the first part of the study. This replication was attempted only for the PSAT/NMSQT Form 2 equating to the SAT Second Old Form.

Two techniques were used to evaluate the results of the various methods. First, graphical comparisons are presented to give an overview of the relative agreement of each traditional method with the IRT method or methods. Second,

Group	PSAT/NMSQT Form 1 Unique Items n=20	PSAT/NMSQT Form 1 - SAT First Old Form Common Items n=22	PSAT/NMSQT Form 1 - SAT Second Old Form Common Items n=23	PSAT/NMSQT Form 2 Unique Items n=19	PSAT/NMSQT Form 2 - SAT First Old Form Common Items n=23	PSAT/NMSQT Form 2 - SAT Second Old Form Common Items n=23	SAT First Old Form Unique Items n=40	SAT Second Old Form Unique Items n=39
PSAT/ NMSQT Form 1	X	X	X	Not Reached	Not Reached	Not Reached	Not Reached	Not Reached
PSAT/ NMSQT Form 2	Not Reached	Not Reached	Not Reached	X	X	X	Not Reached	Not Reached
SAT First Old Form	Not Reached	X	Not Reached	Not Reached	X	Not Reached	X	Not Reached
SAT Second Old Form	Not Reached	Not Reached	X	Not Reached	Not Reached	X	Not Reached	X

Figure 3: Calibration Plan for Direct IRT Equating of PSAT/NMSQT Form 1 Verbal Section to PSAT/NMSQT Form 2 Verbal Section. The entire matrix represents a single calibration run. Crosses indicate items that examinee groups were actually exposed to. Each PSAT/NMSQT and SAT sample contains approximately 2,000 cases.

discrepancy indices (cf. Marco et al, 1979 and Appendix B) for the total score distribution and three regions (upper 20, middle 60, and lower 20 percent of this distribution) are provided as a numerical indication of differences across methods. The discrepancy index described by Marco et al (1979), is simply a weighted (weighted by the frequency of the equating samples) mean-squared difference between an estimated score and a criterion score. Since this study is concerned with agreement with, rather than performance against, a criterion, the discrepancy index is here better thought of as an index of agreement. It is thus a weighted mean-squared difference between scaled scores estimated by each of the traditional methods compared to those estimated by the IRT method. Details for calculating the discrepancy index are given in Appendix B.

RESULTS

The results of the first part of the study, which involved the comparison of IRT estimated true formula score equating with three traditional methods, linear, equipercentile and frequency estimation equipercentile, for the four PSAT/NMSQT, SAT pairings are summarized in Tables 1-8 and Figures 4-7. Raw score to scale score transformations for each equating method applied to each PSAT/NMSQT, SAT pairing are given in Tables 1-4. The information contained in these tables is also presented graphically in Figures 4-7. Each figure contains three plots comparing the traditional equating methods with the IRT method. Tables 5-8 contain summary data and discrepany indices computed as a means for comparing the traditional equating methods with the IRT method. Each table contains data for a single PSAT/NMSQT, SAT pairing.

Examination of the information contained in Table 1 and illustrated in Figure 4, indicates close agreement of all three traditional equating methods with the IRT method for the PSAT/NMSQT Form 1 Math, SAT First Old Form pairing. The IRT method tended to yield slightly higher scaled scores than either the linear or traditional curvilinear methods at the extremes of the score scale. The method that appears to agree most closely with the IRT method is the frequency estimation equipercentile method.

Table 2 and Figure 5 contain information pertaining to the PSAT/NMSQT Form 1 Math, SAT Second Old Form pairing. Again, close agreement was found among the raw to scale conversions for all three traditional methods, compared to the IRT method. The IRT method tended to yield slightly lower scaled scores than any of the three traditional methods. The procedure that appears to agree most closely with the IRT equating is again the frequency estimation equipercentile method, although the equipercentile agrees more closely for the upper and lower ends of the score range.

The results of the PSAT/NMSQT Form 2 Verbal, SAT First Old Form equating are presented in Table 3 and Figure 6. It can be seen, from examination of these data, that the IRT equating method yielded higher scaled scores than the linear method, particularly at the extremes of the score scale. The traditional method that agrees most closely with the IRT method appears to be the equipercentile.

Table 4 and Figure 7 contain the raw to scale conversions resulting from application of the four equating methods to the PSAT/NMSQT Form 2 Verbal, SAT Second Old Form pairing. For this equating, the IRT method tended to yield scaled scores that agreed quite well with those obtained by the traditional methods, with the exception of linear conversions at the upper end of the score scale. The method that appears to agree most closely with the IRT

Table 1

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 1 MATH TO SAT FIRST OLD FORM

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
50	2	744.32	766.42	774.7	785.8
49	3	734.67	745.42	753.4	764.2
48	0	725.03	734.84	742.1	746.8
47	0	715.38	724.26	736.4	732.1
46	4	705.73	713.69	723.1	718.8
45	7	696.08	703.11	707.4	705.3
44	8	686.43	693.10	694.9	694.3
43	4	676.78	684.21	684.8	682.6
42	10	667.13	675.31	675.1	671.3
41	18	657.48	663.92	662.6	660.1
40	22	647.83	651.46	646.9	649.1
39	20	638.19	641.17	634.0	638.4
38	15	628.53	630.88	625.8	627.8
37	19	618.88	619.69	617.1	617.4
36	36	609.23	607.86	606.5	607.1
35	44	599.58	596.04	595.5	597.0
34	24	589.93	587.08	587.2	587.0
33	36	580.29	578.25	579.3	577.0
32	39	570.64	569.11	569.0	567.1
31	44	560.99	557.76	557.4	557.3
30	37	551.34	548.01	546.8	547.5
29	31	541.69	540.95	538.0	537.7
28	49	532.04	532.74	528.2	527.9
27	56	522.39	521.06	518.2	518.1
26	65	512.74	509.55	509.4	508.3
25	55	503.09	500.16	500.8	498.6
24	45	493.44	490.77	492.7	488.9
23	63	483.79	482.73	483.7	479.2
22	54	474.14	475.26	473.7	469.6
21	65	464.49	467.47	463.9	460.0
20	59	454.84	458.24	454.5	450.4
19	55	445.20	448.28	443.6	441.0
18	85	435.55	434.65	433.0	431.5
17	78	425.90	422.96	422.3	422.2
16	80	416.25	413.93	410.6	412.9
15	38	406.60	404.92	401.7	403.6
14	56	396.95	397.26	393.9	394.5
13	75	387.30	389.60	384.3	385.4
12	84	377.65	381.19	373.6	376.3
11	54	368.00	369.55	363.0	367.4
10	54	358.35	358.42	354.7	358.5
9	55	348.70	349.34	346.7	349.8
8	50	339.05	339.60	337.2	341.1
7	65	329.40	329.03	326.4	332.4
6	29	319.75	318.68	317.9	323.8
5	52	310.11	309.45	310.6	315.2
4	41	300.46	300.21	301.8	306.6
3	30	290.81	290.29	292.3	298.0

Table 1 (cont.)

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 1 MATH TO SAT FIRST OLD FORM (CON'T)

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
2	12	281.16	279.99	282.0	289.2
1	26	271.51	269.70	274.1	280.2
0	21	261.86	258.78	261.6	271.0
-1	9	252.21	248.38	241.4	261.4
-2	7	242.56	241.51	230.4	251.5
-3	3	232.91	234.94	223.4	241.2
-4	3	223.26	228.50	217.0	230.3
-5	3	213.61	222.06	205.1	218.9
-6	1	203.96	205.77	196.9	208.1

Table 2

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 1 MATH TO SAT SECOND OLD FORM

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
49	2	743.40	751.06	738.4	753.2
48	1	733.67	734.48	731.2	741.8
47	4	723.93	726.30	727.2	730.9
46	3	714.20	718.12	721.6	720.2
45	9	704.47	707.23	712.3	709.6
44	6	694.73	695.44	699.6	699.0
43	7	685.00	684.11	692.4	688.3
42	12	675.27	673.79	680.2	677.4
41	12	665.53	664.06	667.5	666.5
40	11	655.80	655.80	655.0	655.4
39	22	646.06	647.53	643.0	644.4
38	15	636.33	637.38	634.0	633.3
37	21	626.60	626.09	624.7	622.3
36	28	616.86	614.80	614.5	611.5
35	37	607.13	604.17	605.4	600.7
34	25	597.40	593.58	597.5	590.1
33	30	587.66	582.36	589.9	579.6
32	40	577.93	570.00	579.3	569.2
31	60	568.20	558.10	566.5	559.0
30	45	558.46	547.88	553.9	548.9
29	42	548.73	538.87	544.9	538.9
28	41	539.00	531.62	537.1	529.0
27	57	529.26	524.38	526.9	519.3
26	59	519.53	516.24	516.3	509.7
25	41	509.80	507.94	507.6	500.2
24	44	500.06	499.59	498.8	490.7
23	45	490.33	490.43	488.8	481.4
22	70	480.60	481.28	478.9	472.1
21	66	470.86	472.10	468.8	462.8
20	52	461.13	462.85	459.2	453.6
19	65	451.39	453.52	451.2	444.5
18	73	441.66	442.91	442.9	435.4
17	72	431.93	432.87	432.9	426.4
16	56	422.19	423.68	421.9	417.4
15	41	412.46	414.49	413.0	408.4
14	76	402.73	404.82	405.3	399.4
13	71	392.99	395.09	395.3	390.5
12	74	383.26	385.27	384.6	381.6
11	61	373.53	375.35	375.6	372.7
10	49	363.79	364.86	367.8	363.8
9	62	354.06	352.43	358.5	354.9
8	74	344.33	341.16	347.2	345.9
7	64	334.59	331.03	336.2	337.0
6	35	324.86	322.56	327.3	327.9
5	39	315.13	314.83	318.8	318.8
4	48	305.39	307.10	308.9	309.7
3	44	295.66	296.29	298.2	300.4
2	20	285.93	285.29	287.4	290.9
1	16	276.19	276.37	278.8	281.3

Table 2 (cont.)

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM I MATH TO SAT SECOND OLD FORM (CON'T)

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUIPERCENTILE	FREQ EST EQUIPERCENTILE	IRT
0	19	266.46	267.44	266.8	271.6
-1	11	256.72	259.19	253.1	261.8
-2	8	246.99	250.99	245.7	252.1
-3	3	237.26	242.77	238.1	242.4
-4	8	227.52	234.55	228.0	233.0
-5	3	217.79	225.50	220.5	223.4
-6	1	208.06	216.27	212.0	213.0

Table 3

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT FIRST OLD FORM

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
64	1	729.56	744.06	766.6	769.7
63	1	720.95	735.61	741.4	756.6
62	0	712.34	727.15	724.1	744.9
61	1	703.74	718.70	716.0	734.0
60	2	695.13	711.43	708.0	723.5
59	0	686.52	704.16	699.0	713.3
58	1	677.91	696.89	687.9	703.2
57	0	669.31	689.62	681.5	693.2
56	8	660.70	679.28	674.4	683.2
55	6	652.09	665.30	665.1	673.2
54	7	643.48	653.51	654.1	663.2
53	12	634.88	642.41	646.4	653.1
52	5	626.27	633.08	639.8	643.0
51	8	617.66	623.75	627.4	632.9
50	15	609.05	614.43	617.1	622.8
49	19	600.45	604.15	608.0	612.6
48	27	591.84	593.72	595.4	602.4
47	9	583.23	583.45	585.3	592.3
46	17	574.62	576.06	578.9	582.3
45	16	566.02	568.68	569.5	572.3
44	20	557.41	560.76	557.2	562.5
43	30	548.80	550.22	544.6	552.7
42	21	540.19	539.63	538.0	543.1
41	43	531.59	527.97	531.9	533.5
40	36	522.98	516.31	522.5	524.1
39	39	514.37	506.24	511.6	514.8
38	49	505.76	497.28	502.1	505.6
37	29	497.16	488.32	494.6	496.6
36	51	488.55	480.62	486.5	487.6
35	45	479.94	474.40	476.8	478.8
34	46	471.33	468.17	467.9	470.1
33	60	462.73	461.94	459.1	461.5
32	29	454.12	452.45	451.5	453.0
31	82	445.51	442.07	443.5	444.6
30	83	436.90	432.19	434.6	436.3
29	67	428.30	423.29	425.5	428.1
28	72	419.69	415.05	416.1	419.9
27	34	411.08	408.47	408.3	411.9
26	67	402.47	401.88	401.2	403.9
25	73	393.87	395.30	392.6	395.9
24	55	385.26	386.89	383.7	388.0
23	76	376.65	377.84	372.7	380.0
22	38	368.04	368.80	364.8	372.1
21	79	359.44	359.52	358.2	364.2
20	72	350.83	349.89	349.6	356.2
19	58	342.22	340.26	339.1	348.3
18	65	333.61	332.09	330.1	340.3
17	33	325.01	325.35	324.2	332.3

Table 3 (cont.)

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT FIRST OLD FORM (CON'T)

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
16	56	316.40	313.62	318.1	324.2
15	48	307.79	309.12	308.3	316.1
14	42	299.18	300.48	298.8	308.0
13	34	290.58	293.85	289.9	299.9
12	15	281.97	287.23	283.7	291.7
11	36	273.36	280.42	276.4	283.4
10	31	264.75	271.98	265.7	275.1
9	27	256.15	263.54	257.7	266.8
8	25	247.54	255.23	250.9	258.3
7	10	238.93	247.21	245.4	249.8
6	17	230.32	239.18	238.2	241.2
5	16	221.72	229.86	226.9	232.5
4	12	213.11	217.54	218.4	223.7
3	8	204.50	206.97	212.3	214.8
2	2	195.89	198.20	208.7	205.7
1	4	187.29	192.00	202.3	196.5
0	3	178.68	185.79	190.7	187.2
-1	4	170.07	179.68	180.9	177.5
-2	1	161.46	173.80	177.1	167.6
-3	0	152.86	167.92	170.4	157.0
-4	2	144.25	153.83	161.0	145.1

Table 4

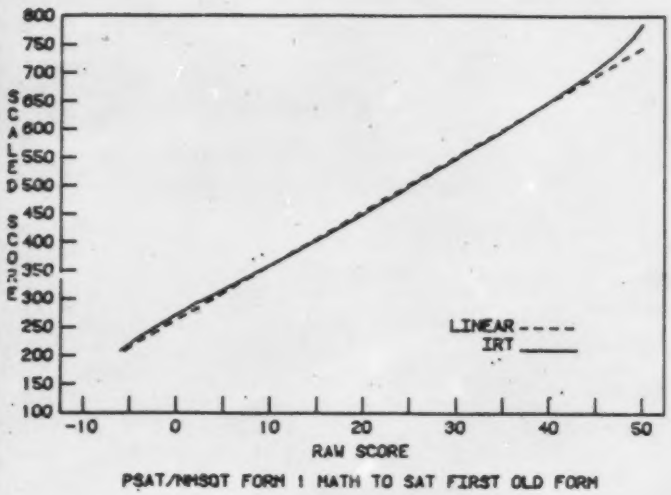
A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT SECOND OLD FORM

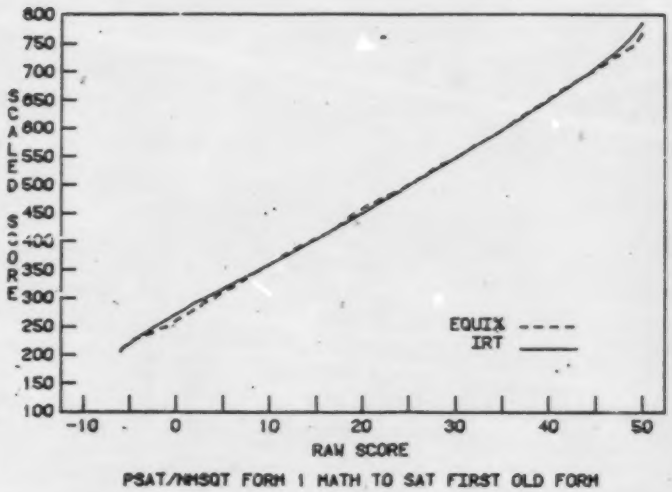
RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
63	2	738.75	749.85	779.4	765.4
62	0	729.93	742.64	773.1	753.1
61	1	721.10	735.42	753.4	741.3
60	1	712.27	728.21	730.9	729.9
59	3	703.45	717.41	714.1	719.9
58	4	694.62	705.41	695.4	708.1
57	1	685.79	693.41	685.5	697.5
56	5	676.97	684.50	679.0	687.1
55	2	668.14	678.69	671.8	676.9
54	7	659.32	672.89	660.8	666.7
53	9	650.49	660.05	651.1	656.6
52	5	641.66	645.89	644.3	646.5
51	8	632.84	632.80	636.7	636.4
50	8	624.10	619.72	628.3	626.3
49	14	615.18	609.19	619.6	616.2
48	15	606.36	601.21	609.0	606.1
47	7	597.53	593.23	600.6	596.1
46	18	588.71	584.99	591.5	586.2
45	32	579.88	576.46	582.1	576.3
44	23	571.05	567.93	572.6	566.6
43	41	562.23	556.79	563.2	557.0
42	13	553.40	545.52	555.3	547.5
41	26	544.58	536.11	547.0	538.1
40	35	535.75	526.70	536.7	528.9
39	38	526.92	518.31	524.8	519.8
38	41	518.10	511.07	513.4	510.7
37	23	509.27	503.84	506.1	501.9
36	59	500.44	496.37	499.0	493.1
35	55	491.62	488.40	489.8	484.4
34	55	482.79	480.42	480.2	475.7
33	70	473.97	472.24	471.3	467.2
32	34	465.14	463.88	464.7	458.7
31	70	456.31	455.52	458.5	450.3
30	67	447.49	447.15	448.9	441.8
29	58	438.66	438.75	439.2	433.4
28	73	429.83	430.35	429.8	425.1
27	38	421.01	421.42	422.1	416.7
26	85	412.18	412.37	414.3	408.3
25	61	403.36	403.33	404.7	399.9
24	75	394.53	394.29	395.5	391.5
23	77	385.70	385.38	386.2	383.0
22	37	376.88	378.00	378.6	374.6
21	64	368.05	370.62	371.5	366.1
20	71	359.22	363.16	362.8	357.6
19	74	350.40	353.21	353.3	349.2
18	52	341.57	343.26	342.2	340.7
17	23	332.75	334.60	334.2	332.3
16	41	323.92	326.91	326.4	323.9

Table 4 (cont.)

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT SECOND OLD FORM (CON'T)

RAW SCORE	FREQ	ESTIMATED SCALED SCORE			
		LINEAR	EQUI%	FREQ EST EQUI%	IRT
15	54	315.09	318.63	316.7	315.5
14	56	306.27	305.37	306.4	307.1
13	42	297.44	295.66	296.9	298.8
12	20	288.62	288.74	290.3	290.4
11	39	279.79	281.81	282.8	282.0
10	26	270.96	273.31	270.4	273.7
9	26	262.14	264.48	261.9	265.3
8	22	253.31	254.67	253.1	256.8
7	15	244.48	243.50	246.2	248.4
6	24	235.66	232.20	239.8	239.9
5	16	226.83	220.01	231.0	231.4
4	11	218.01	210.50	222.7	222.9
3	4	209.18	202.50	215.1	214.3
2	4	200.35	196.88	210.7	205.8
1	6	191.53	191.26	205.6	197.3
0	3	182.70	185.64	192.7	188.9
-1	3	173.87	180.14	177.1	180.4
-2	2	165.05	175.25	167.0	171.9
-3	1	156.22	170.36	157.4	163.4
-4	2	147.40	165.47	149.0	154.7





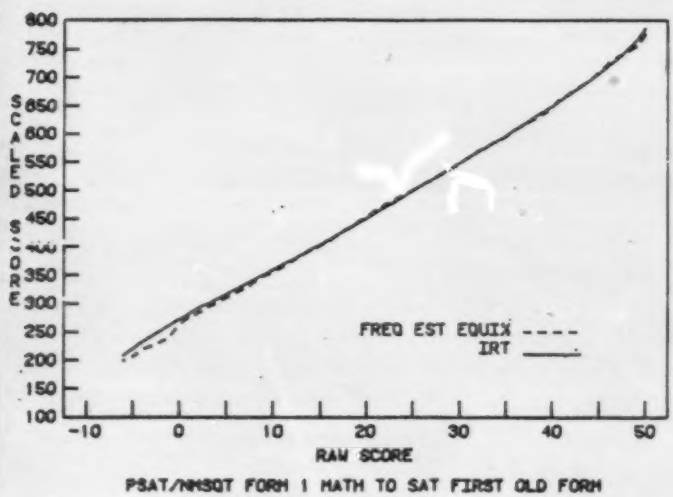
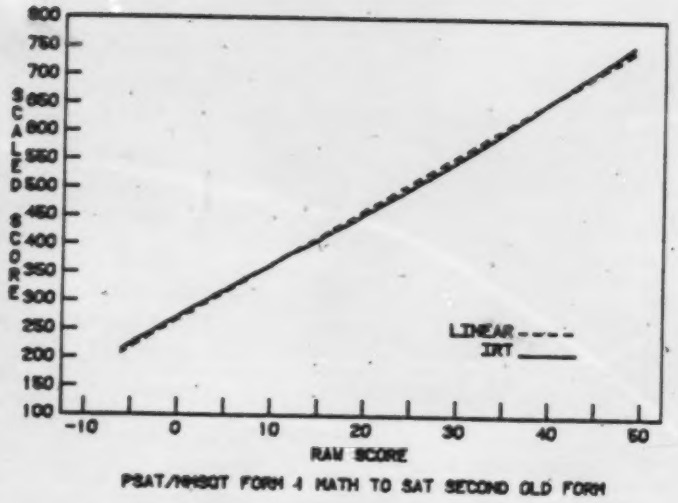
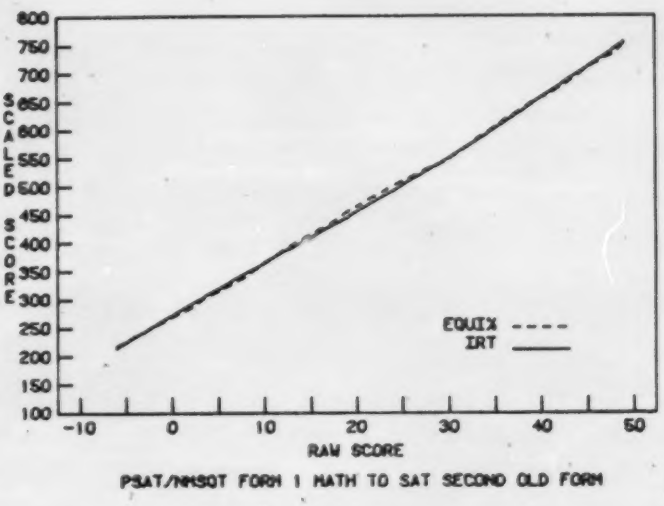


Figure 4: Comparison of Raw Score to Scale Score Conversions obtained by Traditional and IRT Methods for PSAT/NMSQT Form 1 Math - SAT First Old Form Equatings.





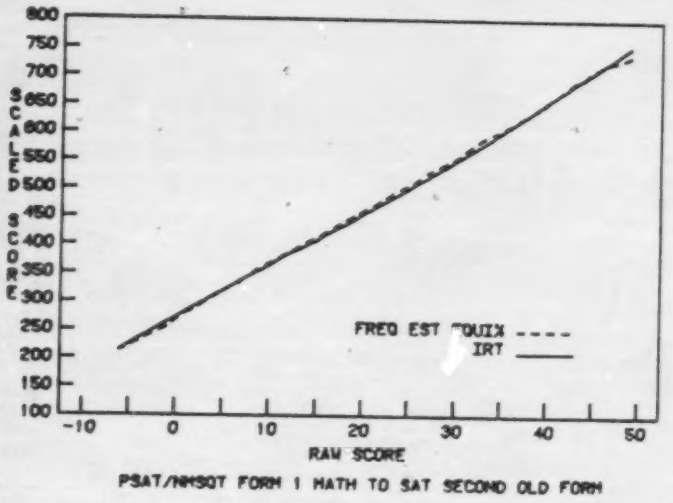
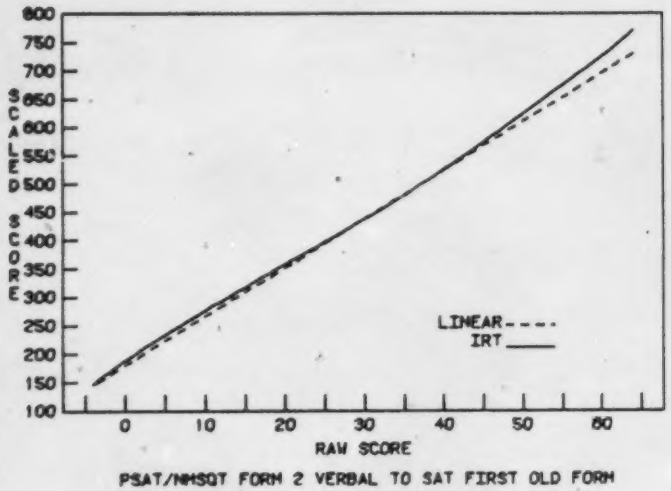
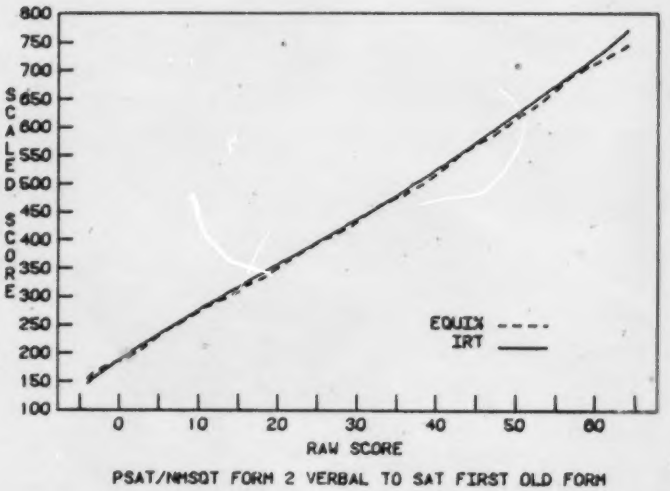


Figure 5: Comparison of Raw Score to Scale Score Conversions obtained by Traditional and IRT Methods for PSAT/NMSQT Form 1 Math - SAT Second Old Form Equatings.





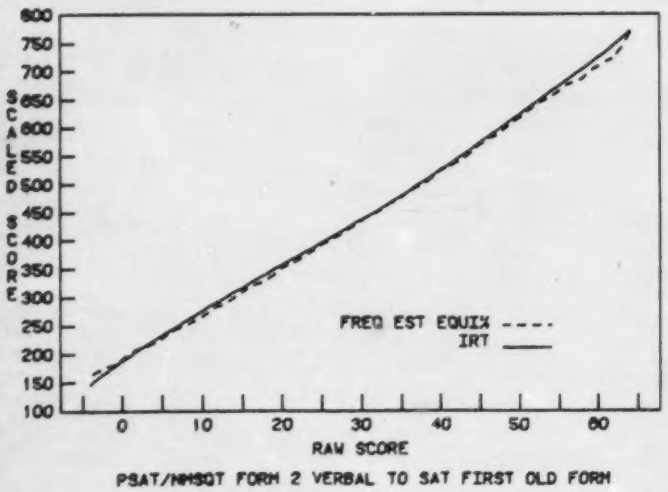
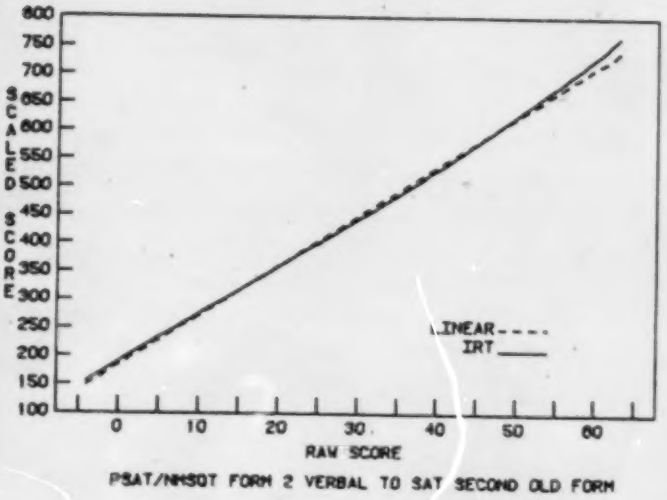
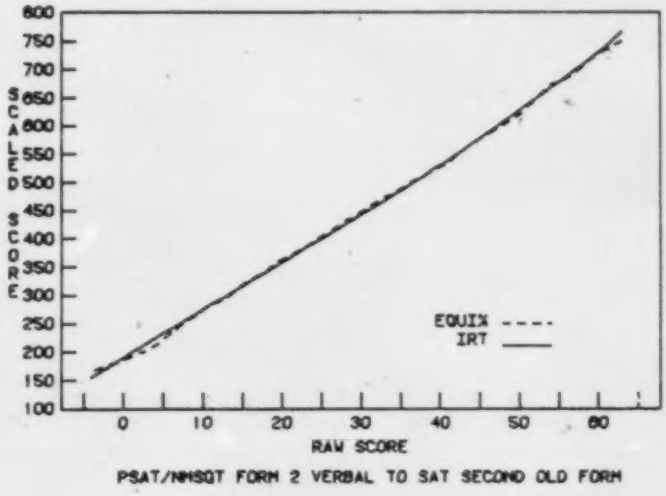


Figure 6: Comparison of Raw Score to Scale Score Conversions obtained by Traditional and IRT Methods for PSAT/NMSQT Form 2 Verbal - SAT First Old Form Equatings.





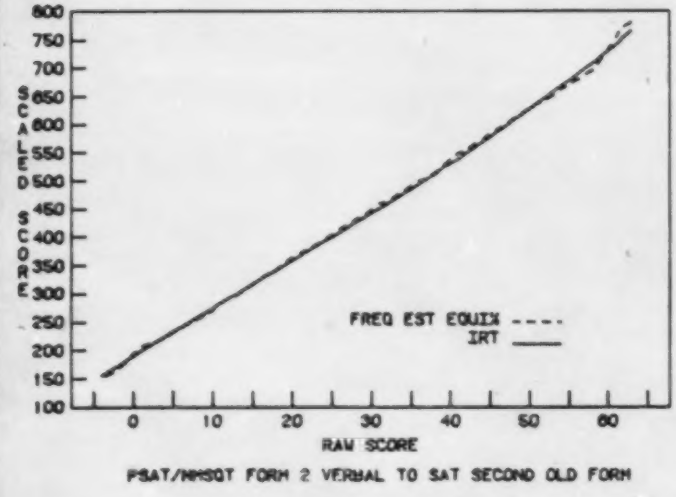


Figure 7: Comparison of Raw Score to Scale Conversions obtained by Traditional and IRT Methods for PSAT/NMSQT Form 2 Verbal - SAT Second Old Form Equatings.

Table 5

Summary of Discrepancy Indices for Equating Methods
PSAT/NMSQT Form 1 Mathematical Section - SAT First Old Form
Total Score Distribution and Three Subdivisions

		Equating Methods			
		IRT	Linear	Equipercentile	Frequency Estimation Equipercentile
Total Score	Scaled Score Mean	447.37	448.61	448.52	446.43
	Scaled Score Standard Deviation	103.05	104.62	104.79	105.26
Total Score Distribution	Total Error		19.63	19.77	15.41
	Bias		1.24	1.15	-.94
	Standard Deviation of Difference		4.25	4.30	3.81
Upper 20% of Distribution	Total Error		28.50	8.27	5.04
	Bias		.91	.80	-.13
	Standard Deviation of Difference		5.26	2.76	2.24
Middle 60% of Distribution	Total Error		13.36	16.73	7.36
	Bias		3.38	3.40	.52
	Standard Deviation of Difference		1.39	2.27	2.66
Lower 20% of Distribution	Total Error		29.61	39.82	49.13
	Bias		-4.75	-5.17	-6.02
	Standard Deviation of Difference		2.65	3.62	3.59

Table 6

Summary of Discrepancy Indices for Equating Methods
PSAT/NMSQT Form 1 Mathematical Section - SAT Second Old Form
Total Score Distribution and Three Subdivisions

		Equating Methods			
		IRT	Linear	Equipercentile	Frequency Estimation Equipercentile
Total Score	Scaled Score Mean	446.82	450.67	449.61	451.04
	Scaled Score Standard Deviation	103.41	106.47	104.93	105.25
Total Score Distribution	Total Error		39.92	31.29	30.97
	Bias		3.85	2.80	4.22
	Standard Deviation of Difference		5.01	4.84	3.62
Upper 20% of Distribution	Total Error		49.98	7.20	39.17
	Bias		5.42	.91	4.89
	Standard Deviation of Difference		4.54	2.52	3.90
Middle 60% of Distribution	Total Error		44.96	42.46	36.17
	Bias		5.74	5.67	5.80
	Standard Deviation of Difference		3.47	3.20	1.58
Lower 20% of Distribution	Total Error		14.32	20.61	6.74
	Bias		-3.54	-4.24	-1.33
	Standard Deviation of Difference		1.33	1.63	2.23

Table 7

Summary of Discrepancy Indices for Equating Methods
PSAT/NMSQT Form 2 Verbal Section - SAT First Old Form
Total Score Distribution and Three Subdivisions

		Equating Methods			
		IRT	Linear	Equipercentile	Frequency Estimation Equipercentile
Total Score	Scaled Score Mean	414.98	410.77	410.31	410.12
	Scaled Score Standard Deviation	100.25	101.05	99.64	101.21
Total Score Distribution	Total Error		42.04	29.46	32.75
	Bias		-4.21	-4.67	-4.86
	Standard Deviation of Difference		4.93	2.77	3.02
Upper 20% of Distribution	Total Error		89.37	55.00	27.88
	Bias		-6.37	-6.86	-4.61
	Standard Deviation of Difference		6.99	2.82	2.57
Middle 60% of Distribution	Total Error		11.07	21.88	25.91
	Bias		-1.83	-3.95	-4.28
	Standard Deviation of Difference		2.78	2.51	2.75
Lower 20% of Distribution	Total Error		90.40	27.35	58.76
	Bias		-9.42	-4.70	-6.90
	Standard Deviation of Difference		1.27	2.28	3.35

Table 8

Summary of Discrepancy Indices for Equating Methods
PSAT/NMSQT Form 2 Verbal Section - SAT Second Old Form
Total Score Distribution and Three Subdivisions

		Equating Methods			
		IRT	Linear	Equipercentile	Frequency Estimation Equipercentile
Total Score	Scaled Score Mean	414.60	417.25	416.54	418.30
	Scaled Score Standard Deviation	101.22	102.97	101.66	102.56
Total Score Distribution	Total Error		22.77	17.16	23.95
	Bias		2.66	1.95	3.71
	Standard Deviation of Difference		3.96	3.66	3.19
Upper 20% of Distribution	Total Error		42.14	8.15	35.08
	Bias		3.29	-1.09	4.26
	Standard Deviation of Difference		5.60	2.64	4.11
Middle 60% of Distribution	Total Error		21.07	18.46	26.20
	Bias		4.01	4.17	4.89
	Standard Deviation of Difference		2.22	1.05	1.52
Lower 20% of Distribution	Total Error		8.74	21.99	5.26
	Bias		-2.44	-2.27	-.74
	Standard Deviation of Difference		1.67	4.11	2.17

method is again the equipercentile.

Further insight into the differential effects of the equating methods on the raw to scale score transformations can be gained by examination of the discrepancy indices computed for the total score distribution and for three segments of this distribution. It should be re-emphasized at this point that the traditional equating methods are being assessed in terms of their agreement with the IRT method. Therefore the term "total error" should be thought of as a measure of agreement, not necessarily as a measure of error.

Table 5 presents summary data and discrepancy indices for the PSAT/NMSQT Form 1 Math, SAT First Old Form equating. Examination of these data indicates that the IRT method yielded a slightly lower estimate of the mean than the linear and equipercentile method and a slightly higher estimate than the frequency estimation equipercentile method. The IRT method produced slightly smaller estimates of the standard deviation than any of the traditional methods. Examination of the discrepancy indices for the total score distribution and for the three segments of this distribution indicates that most of the discrepancy between the IRT and the linear method occurs at the extremes of the score distribution, i.e., the IRT method agrees much better with both of the curvilinear methods (equipercentile and frequency estimation equipercentile) at the upper 20% of the distribution than it does with the linear method.

The discrepancy indices and summary information for the PSAT/NMSQT Form 1 Math, SAT Second Old Form pairing are given in Table 6. The data indicate that, for the total score distribution, the linear equating method yielded the most discrepant results when compared to the IRT method. This discrepancy can be attributed mostly to disagreement at the upper extreme and middle portion of the score distribution. The IRT method agrees very well with the equiper-

centile method for the upper 20% of the distribution and shows even better agreement with the frequency estimation equipercentile procedure for the lower 20%. The IRT method yielded a slightly lower estimate of the mean and smaller estimate of the standard deviation when compared to the other three equating methods.

Table 7 contains information pertaining to the discrepancy indices and summary statistics computed for the PSAT/NMSQT Form 2 Verbal, SAT First Old Form equating. Examination of the data indicates that the IRT method produced a slightly higher estimate of the mean than the three traditional methods and a slightly smaller estimate of the standard deviation for all the methods, except the equipercentile procedure. As was the case with the previous equatings, the linear method appears to be the most discrepant. It is interesting to note that in this case, although they provide more agreement with the IRT method than the linear method, both curvilinear results are quite discrepant from the IRT results at the extremes of the distribution.

The results of the discrepancy index computations and summary statistics for the PSAT/NMSQT Form 2 Verbal, SAT Second Old Form equating are presented in Table 8. For this equating, the IRT method produced slightly lower estimates of the mean and smaller estimates of the standard deviation than any of the traditional methods. The linear method appears to be the most discrepant for scores in the upper 20% of the distribution. The IRT and equipercentile results show close agreement for this segment of the distribution as do the IRT and frequency estimation equipercentile method for the lower 20% of the distribution.

The results of the second part of the study, which investigated the feasibility of using IRT to equate the PSAT/NMSQT Form 1 and Form 2 Verbal

scores directly, are presented in Tables 9 and 10 and Figure 8. Table 9 contains the raw to scale conversions obtained from the direct PSAT/NMSQT Form 2 to Form 1 equating compared to each of the four previous equatings performed for the PSAT/NMSQT Form 2 Verbal, SAT First Old Form pairing. It should be noted at this point that the calibration design permits the PSAT/NMSQT form to form equating to be carried out in several different ways, e.g., Form 1 could have been equated to Form 2 and both tests placed on scale through the Form 2, SAT Second Old Form relationship. The direction of equating used in the study was chosen to minimize the amount of linear interpolation involved, thus reducing the possibility of the interpolation process contributing to error which might confound the results. The column labeled IRT(2A) in Table 9 contains raw to scale conversions that are the result of placing PSAT/NMSQT Form 2 Verbal scores on the SAT First Old Form scale after equating Form 2 of the PSAT/NMSQT to Form 1. Figure 8 depicts the information given in Table 9 graphically. Each of the four previously performed equatings are compared to the PSAT/NMSQT direct form to form equating. Table 10 contains discrepancy indices computed from a comparison of each of the four previously performed equatings with the direct form to form equating.

Examination of the information contained in Table 9 and illustrated in Figure 8, shows very close agreement between the IRT results obtained from equating the PSAT/NMSQT Form 2 Verbal scores to the SAT First Old Form (labeled IRT) and those obtained from the direct equating of the test to the PSAT/NMSQT Form 1 Verbal scores (labeled IRT(2A)).

Table 10 contains discrepancy index information comparing the IRT(2A) results with those obtained from the three traditional equatings and the IRT equating. The data indicates that the IRT(2A) equating results tended to

Table 9

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT FIRST OLD FORM

RAW SCORE	FREQ	ESTIMATED SCALED SCORE				
		LINEAR	EQUI%	FREQ EST EQUI%	IRT	IRT(2A)
64	1	729.56	744.06	766.6	769.7	770.73
63	1	720.95	735.61	741.4	756.6	757.94
62	0	712.34	727.15	724.1	744.9	746.50
61	1	703.74	718.70	716.0	734.0	735.65
60	2	695.13	711.43	708.0	723.5	725.06
59	0	686.52	704.16	699.0	713.3	714.88
58	1	677.91	696.89	687.9	703.2	704.49
57	0	669.31	689.62	681.5	693.2	694.28
56	8	660.70	679.28	674.4	683.2	684.11
55	6	652.09	665.30	665.1	673.2	673.94
54	7	643.49	653.51	654.1	663.2	663.77
53	12	634.88	642.41	646.4	653.1	653.58
52	5	626.27	633.08	639.8	643.0	643.37
51	8	617.66	623.75	627.4	632.9	633.13
50	15	609.05	614.43	617.1	622.8	622.89
49	12	600.45	604.15	608.0	612.6	612.66
48	27	591.84	593.72	595.4	602.4	602.44
47	9	583.23	583.45	585.3	592.3	592.28
46	17	574.62	576.06	578.9	582.3	582.16
45	16	566.02	568.69	569.5	572.3	572.13
44	20	557.41	560.76	557.2	562.5	562.17
43	30	548.80	550.22	544.6	552.7	552.28
42	21	540.19	539.63	538.0	543.1	542.62
41	43	531.59	527.97	531.9	533.5	532.88
40	36	522.98	516.31	522.5	524.1	523.33
39	37	514.37	506.24	511.6	514.8	513.89
38	49	505.76	497.28	502.1	505.6	504.57
37	29	497.16	489.32	494.6	496.6	495.36
36	51	488.55	480.62	486.5	487.6	486.28
35	45	479.94	474.40	476.8	478.8	477.35
34	46	471.33	468.17	467.9	470.1	468.54
33	60	462.73	461.94	459.1	461.5	459.87
32	29	454.12	452.45	451.5	453.0	451.33
31	82	445.51	442.07	443.5	444.6	442.92
30	83	436.90	432.19	434.6	436.3	434.62
29	67	428.30	423.29	425.5	428.1	426.43
28	72	419.69	415.05	416.1	419.9	418.32
27	34	411.08	408.47	408.3	411.9	410.29
26	67	402.47	401.88	401.2	403.9	402.33
25	73	393.87	395.30	392.6	395.9	394.40
24	55	385.26	386.89	383.7	388.0	386.49
23	76	376.65	377.84	372.7	380.0	378.60
22	38	368.04	368.80	364.8	372.1	370.70
21	79	359.44	359.52	358.2	364.2	362.79
20	72	350.83	349.89	349.6	356.2	354.86
19	58	342.22	340.26	339.1	348.3	346.90
18	65	333.61	332.09	330.1	340.3	338.92
17	33	325.01	325.35	324.2	332.3	330.90

Table 9 (cont.)

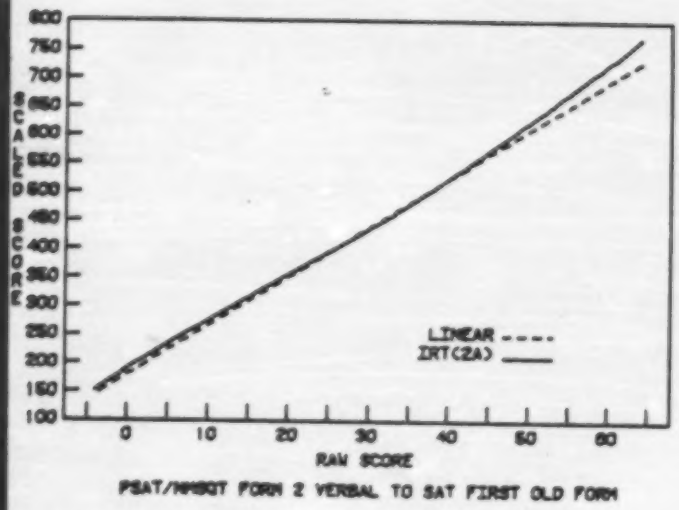
A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT FIRST OLD FORM (CON'T)

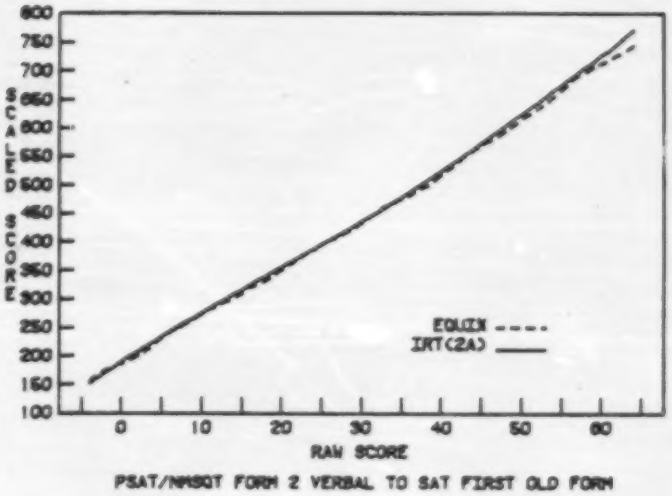
RAW SCORE	FREQ	ESTIMATED SCALED SCORE				
		LINEAR	EQUI%	FREQ EST EQUI%	IRT	IRT(2A)
16	56	316.40	318.62	318.1	324.2	322.85
15	48	307.79	309.12	308.3	316.1	314.78
14	42	299.18	300.48	298.8	308.0	306.67
13	34	290.58	293.85	289.9	299.9	298.53
12	15	281.97	287.23	283.7	291.7	290.37
11	36	273.36	280.42	276.4	283.4	282.18
10	31	264.75	271.98	265.7	275.1	273.94
9	27	256.15	263.54	257.7	266.8	265.68
8	25	247.54	255.23	250.9	259.3	257.37
7	10	238.93	247.21	245.4	249.8	249.02
6	17	230.32	239.19	238.2	241.2	240.62
5	16	221.72	229.86	226.9	232.5	232.17
4	12	213.11	217.54	218.4	223.7	223.63
3	3	204.50	206.97	212.3	214.8	215.02
2	2	195.89	198.20	208.7	205.7	206.32
1	4	187.29	192.00	202.3	196.5	197.50
0	3	178.68	185.79	190.7	187.2	188.52
-1	4	170.07	179.68	180.9	177.5	179.34
-2	1	161.46	173.80	177.1	167.6	169.89
-3	1	152.36	167.92	170.4	157.0	160.06
-4	2	144.25	153.83	161.0	145.1	149.32

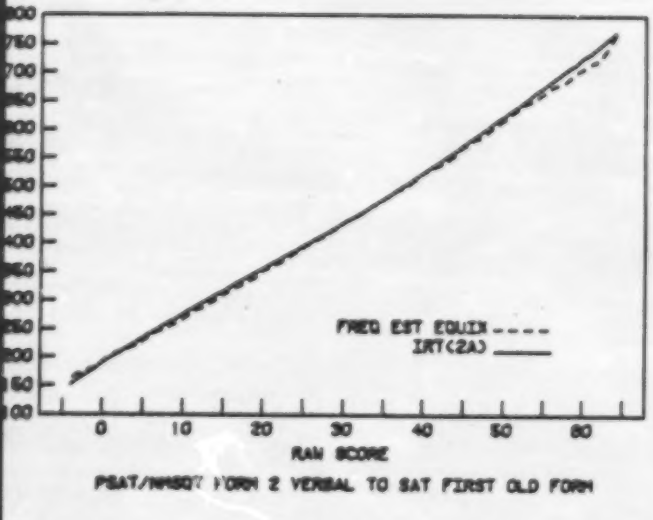
Table 10

Summary of Discrepancy Indices for Equating Methods
PSAT/NMSQT Form 2 Verbal - SAT First Old Form
Total Score Distribution and Three Subdivisions

		Equating Methods				
		IRT (2A)	IRT	Linear	Equipercentile	Frequency Estimation Equipercentile
Total Score	Scaled Score Mean	413.80	414.98	410.77	410.31	410.12
	Scaled Score Standard Deviation	100.46	100.25	101.05	99.64	101.21
Total Score Distribution	Total Error		1.76	37.66	21.23	23.23
	Bias		1.18	-3.04	-3.49	-3.68
	Standard Deviation of Difference		.61	5.33	3.01	3.11
Upper 20% of Distribution	Total Error		.44	92.97	51.31	27.21
	Bias		.35	-6.02	-6.51	-4.27
	Standard Deviation of Difference		.56	7.53	2.98	3.00
Middle 60% of Distribution	Total Error		2.28	8.30	12.64	15.74
	Bias		1.50	-.32	-2.44	-2.78
	Standard Deviation of Difference		2.13	2.86	2.58	2.83
Lower 20% of Distribution	Total Error		1.47	73.02	17.71	42.41
	Bias		.99	-8.43	-3.71	-5.91
	Standard Deviation of Difference		.70	1.39	1.98	2.75







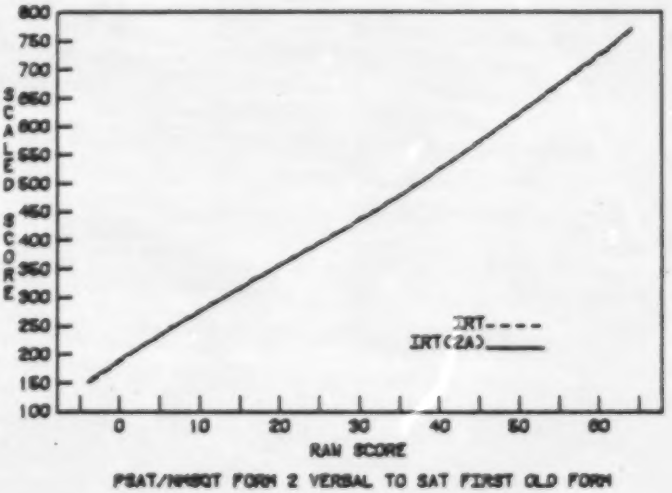


Figure 8: Comparison of Raw Score to Scale Score Conversions obtained by Traditional and two IRT Methods for PSAT/NMSQT Form 2 Verbal - SAT First Old Form Equatings.

yield a slightly lower estimate of the scaled score mean than that obtained from the IRT equating. As noted previously, both curvilinear methods tended to exhibit considerable discrepancy from the IRT method at the extremes of the score distribution. This effect is not quite as pronounced for the IRT(2A) method.

It was noted in an earlier section of this paper that the calibration design used for the PSAT/NMSQT Verbal form to form equating permits the replication of the IRT equating performed for the PSAT/NMSQT Form 2 Verbal, SAT Second Old Form pairing. The results of this equating, designated IRT(2B), are presented in Tables 11 and 12 and Figure 9. Table 11 provides raw to scale conversions for the four previous equatings carried out for this pairing as well as those obtained for the IRT(2B) method. Figure 9 contains a single graph comparing the raw to scale conversions obtained from the IRT and the IRT(2B) methods. Table 12 presents the discrepancy indices computed for the total score distribution and three segments of the score distribution for the IRT-IRT(2B) comparison only.

Examination of the data contained in Tables 11 and 12 and Figure 9 indicates very close agreement between the two IRT methods. It can be seen, from examination of the tabeled data, that the IRT(2B) method tended to yield a very slightly higher estimate of the scaled score mean and slightly smaller estimate of the standard deviation when compared to the IRT method.

CONCLUSIONS

The absence of a true criterion against which to compare the equating methods used in this study somewhat confounds the interpretation of the results. The study assumes that the most appropriate method to use when

Table 11

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT SECOND OLD FORM

RAW SCORE	FREQ	ESTIMATED SCALED SCORE				
		LINEAR	EQUI%	FREQ EST EQUI%	IRT	IRT(29)
63	2	738.75	749.85	779.4	765.4	763.87
62	0	729.93	742.64	773.1	753.1	751.47
61	1	721.10	735.42	753.4	741.3	739.54
60	1	712.27	728.21	730.9	729.9	727.97
59	3	703.45	717.41	714.1	718.9	716.71
58	4	694.62	705.41	695.4	708.1	705.74
57	1	685.79	693.41	685.5	697.5	695.01
56	5	676.97	684.50	679.0	687.1	684.47
55	2	668.14	678.69	671.8	676.9	674.09
54	7	659.32	672.89	660.8	666.7	663.82
53	9	650.49	660.05	651.1	656.6	653.65
52	5	641.66	645.89	644.3	646.5	643.56
51	8	632.84	632.30	636.7	636.4	633.53
50	8	624.10	619.72	628.3	626.3	623.57
49	14	615.18	609.19	619.6	616.2	613.70
48	15	606.36	601.21	609.0	606.1	603.91
47	7	597.53	593.23	600.6	596.1	594.22
46	19	588.71	584.99	591.5	586.2	584.64
45	32	579.88	576.46	582.1	576.3	575.17
44	23	571.05	567.93	572.6	566.6	565.81
43	41	562.23	556.79	563.2	557.0	556.57
42	13	553.40	545.52	555.3	547.5	547.45
41	26	544.58	536.11	547.0	538.1	539.43
40	35	535.75	526.70	536.7	528.9	529.50
39	38	526.92	518.31	524.8	519.8	520.66
38	41	518.10	511.07	513.4	510.7	511.90
37	23	509.27	503.84	506.1	501.9	503.21
36	59	500.44	496.37	499.0	493.1	494.58
35	55	491.62	488.40	489.8	484.4	486.01
34	55	482.79	480.42	480.2	475.7	477.48
33	70	473.97	472.24	471.3	467.2	468.98
32	34	465.14	463.83	464.7	458.7	460.51
31	70	456.31	455.52	458.5	450.3	452.06
30	67	447.49	447.15	448.9	441.8	443.61
29	58	438.66	438.75	439.2	433.4	435.16
28	73	429.83	430.35	429.8	425.1	426.70
27	38	421.01	421.42	422.1	416.7	418.23
26	85	412.18	412.37	414.3	408.3	409.73
25	61	403.36	403.33	404.7	399.9	401.20
24	75	394.53	394.29	395.5	391.5	392.65
23	77	385.70	385.38	386.2	383.0	384.07
22	37	376.88	378.00	378.6	374.6	375.48
21	64	368.05	370.62	371.5	366.1	366.88
20	71	359.22	363.16	362.8	357.6	358.28
19	74	350.40	353.21	353.3	349.2	349.70
18	52	341.57	343.26	342.2	340.7	341.15
17	23	332.75	334.60	334.2	332.3	332.63
16	41	323.92	326.91	326.4	323.9	324.15

Table 11 (cont.)

A COMPARISON OF RAW SCORE TO SCALED SCORE TRANSFORMATIONS
PSAT/NMSQT FORM 2 VERBAL TO SAT SECOND OLD FORM (CON'T)

RAW SCORE	FREQ	ESTIMATED SCALED SCORE				
		LINEAR	EQUI%	FREQ EST EQUI%	IRT	IRT(28)
15	54	315.09	318.63	316.7	315.5	315.71
14	56	306.27	305.37	306.4	307.1	307.31
13	42	297.44	295.66	296.9	298.8	298.95
12	20	288.62	288.74	290.3	290.4	290.62
11	39	279.79	281.81	282.8	282.0	282.30
10	26	270.96	273.31	270.4	273.7	274.00
9	26	262.14	264.48	261.9	265.3	265.70
8	22	253.31	254.67	253.1	256.8	257.40
7	14	244.48	243.50	246.2	248.4	249.09
6	24	235.66	232.20	239.8	239.9	240.78
5	16	226.83	220.01	231.0	231.4	232.45
4	11	218.01	210.50	222.7	222.9	224.12
3	4	209.19	202.50	215.1	214.3	215.79
2	4	200.35	196.98	210.7	205.8	207.47
1	6	191.53	191.26	205.6	197.3	199.14
0	3	182.70	185.64	192.7	188.9	190.80
-1	3	173.87	180.14	177.1	180.4	182.45
-2	2	165.05	175.25	167.0	171.9	174.05
-3	1	156.22	170.36	157.4	163.4	165.55
-4	2	147.40	165.47	149.0	154.7	156.84

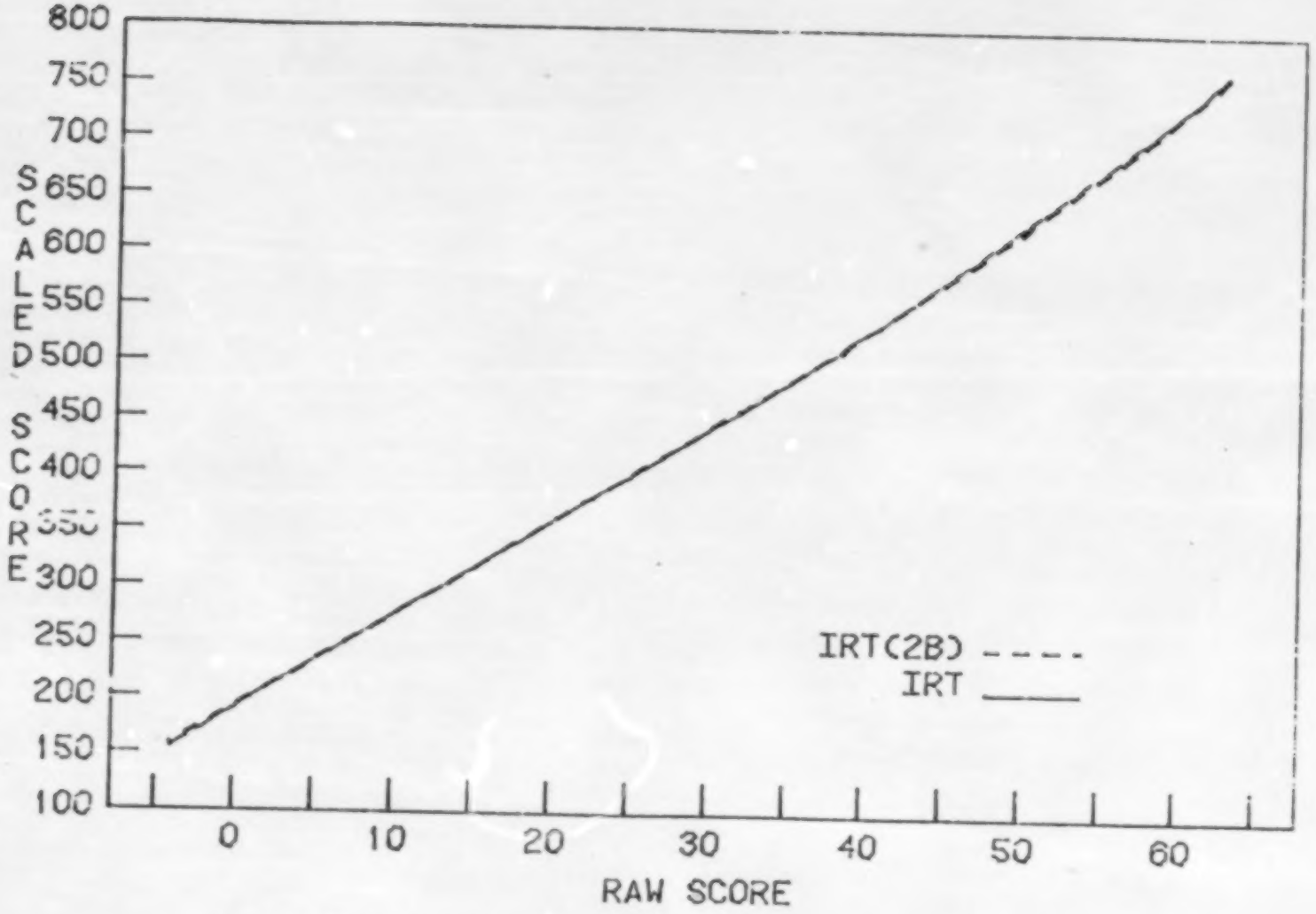


Figure 9: Comparison of Raw Score to Scale Score Conversions obtained by IRT and IRT (2B) Methods for PSAT/NMSQT Form 2 Verbal - SAT Second Old Form Equatings.

Table 12

Summary of Discrepancy Indices for IRT Versus IRT (2B)

PSAT/NMSQT Form 2 Verbal - SAT Second Old Form

Total Score Distribution and Three Subdivisions

Total Score Distribution

Equating Method	Scaled Score Mean	Scaled Score Standard Deviation	Total Error	Bias	Standard Deviation of the Difference
IRT (2B)	415.37	100.90	1.64	.78	1.02
IRT	414.60	101.22			

Upper 20% of Distribution

Equating Method	Total Error	Bias	Standard Deviation of the Difference
IRT (2B)	2.21	-.50	1.40

Middle 60% of Distribution

Equating Method	Total Error	Bias	Standard Deviation of the Difference
IRT (2B)	1.83	1.26	.50

Lower 20% of Distribution

Equating Method	Total Error	Bias	Standard Deviation of the Difference
IRT (2B)	.46	.49	.46

equating dissimilar tests given to non-randomly equivalent groups is an IRT method employing the three-parameter logistic model. There has been substantial support, in the literature cited previously, for this assumption. However, conclusions drawn from the study, based on the assumption that the IRT method is most appropriate, must be considered tentative and subject to verification through replication. A modest attempt at replication was made by equating the PSAT/NMSQT Form 2 Verbal scores to the SAT Second Old Form using item parameter estimates obtained from the two different calibration designs. The comparability of the results of these two equatings lends some credence to the comparisons made between the IRT and traditional methods.

The most notable aspect of the results obtained from the first part of the study, which was designed principally to compare the IRT method to the traditional methods, was the marked agreement found among the four procedures. It appears that all the methods perform fairly similarly for the major portion of the score reporting range, with departures occuring mostly at the extremes of the distribution. As expected, the traditional curvilinear methods agree more closely with the IRT method than does the linear method for these portions of the score scale.

These results run somewhat counter to those suggested by previous research. Previous research involving the equating of tests at different levels of difficulty given to non-randomly equivalent groups suggests that the three-parameter IRT model should work effectively, whereas the traditional methods may not (see Slinde and Linn, 1977; Lord, 1975, 1977). Hence the expected results would be that the traditional methods would not closely coincide with the IRT results, nor would the linear and traditional equipercentile be expected to coincide, simply because the differences in difficulty of the

two forms would force a curvilinear relationship. The unexpected agreement across methods in this study may be partially explained by the fact that the distributions of scores on the two tests were quite similar in shape. In fact, this was somewhat to be expected, given that the tests are constructed to be appropriate for the ability levels of the populations they were administered to. The use of the word populations is important in clarifying the differences between the results of this study and previous research. An underlying assumption of the previous research is that the groups taking the forms are non-randomly equivalent groups from the same population. The differences in difficulty between the two forms is hence sufficient to cause a curvilinear relationship and at the same time theoretically necessitate a true score (IRT) equating method (see Lord, 1980). In this study, while the forms do essentially differ in difficulty, they are constructed to yield the same sort of distribution for the two non-randomly equivalent groups. Thus, the linear and curvilinear methods closely coincide, although the same theoretical argument (Lord, 1980) pertaining to the equating of raw scores on tests of unequal difficulty, would suggest use of a true score or IRT equating method.

One can only conclude, from the results of the first part of the study, that the PSAT/NMSQT, SAT equating is essentially linear, at least for the middle portion of the score reporting range. The fact that the linear method differs from the curvilinear methods at the extremes of the score distribution is evidence that this method, although a good approximation to the curvilinear methods, is not quite appropriate for extreme scores.

Although the traditional curvilinear methods agreed more closely with the IRT method than did the linear procedures at the extremes of the score scale, some discrepancy is apparent. These discrepancies are most probably due to

the fact that the stability of the traditional methods is effected by the scarcity of data at those extremes. Because it is possible to determine the relationship of true scores on two forms of a test for any given θ, regardless of whether it is actually observed, IRT methods are not effected by a lack of data at the upper end of the distribution. This is not true, however, for below chance score level conversions. As explained previously, the three-parameter logistic model does not provide this relationship directly and some method of inferring it must be developed. Therefore, it is difficult to arrive at conclusions indicating that any of the equating methods evaluated provide more appropriate transformations for scores at the extreme low ends of the score scale.

The results of the second part of the study, which investigated the feasibility of using IRT methods to equate the two forms of the PSAT/NMSQT directly were encouraging. Very little difference was found between the scaled scores obtained from the direct form to form IRT equating and those obtained from the IRT equating of the PSAT/NMSQT Form 2 Verbal scores to the SAT First Old Form. This offers some support for the feasibility of using IRT methods for form to form equating of the PSAT/NMSQT.

The fact that the form to form equating appears feasible is important for the following reason. When the two forms of the test are equated separately to the same old SAT form, it is seldom the case that the maximum raw scores on the two forms will be transformed to the same scaled scores. This situation could potentially cause some unfairness to candidates taking the form of the test which yields a lower maximum raw score-scaled score conversion. Typically, scores in the upper region of one of the forms are adjusted slightly such that maximum raw scores on the two forms are transformed to the same scaled score.

This adjustment introduces an unknown degree of error into the equating of the scores at the upper end of the score scale. However, if the two forms are equated directly using the calibration design described in the second part of this study and then placed on scale through their relationship to the same old SAT form, the maximum raw score on both forms will convert to the same scaled score, thus eliminating the necessity of an adjustment to scores in the upper end of the score range.

To summarize, results of the study indicate that traditional linear, equipercentile, frequency estimation equipercentile, and IRT equating using the three-parameter logistic model, provide comparable results for the major portion of the score reporting range even though non-parallel tests given to non-randomly equivalent groups were equated. Where the methods fail to coincide (at the upper end of the distribution), the IRT method is assumed to provide the most appropriate conversions. In addition, a unique application of IRT methods, that of equating non-parallel tests given to non-randomly equivalent groups in the absence of a set of common items or anchor test, has been shown to be feasible.

REFERENCES

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Kolen, M.J. Comparisons of traditional and item response theory methods for equating tests. Journal of Educational Measurement. 1981, 18, 1-11.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F.M. A survey of equating methods based on item characteristic theory. Research Bulletin 75-13. Princeton, NJ: Educational Testing Service, 1975.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Marco, G.L., Petersen, N.S. & Stewart, E.E. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, 1979.

Petersen, N.S., Cook, L.L., and Stocking, M.S. Scale drift: A comparative study of IRT versus linear equating methods. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Slinde, J.A. & Linn, R.L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement 1977, 14, 23-32.

Slinde, J.A. & Linn, R.L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J.A. & Linn, R.L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Wood, R L , & Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, NJ: Educational Testing Service, 1976.

Wood, R L, Wingersky, M.S., & Lord, F.M. LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.